Analysis of Multistate Autoregressive Models

Jie Ding, Shahin Shahrampour, Kathryn Heal, and Vahid Tarokh

***Abstract*—In this paper, we consider the inference problem for a wide class of time-series models, referred to as multistate autoregressive models. The time series that we consider are composed of multiple epochs, each modeled by an autoregressive process. The number of epochs is unknown, and the transitions of states follow a Markov process of an unknown order. We propose an inference strategy that enables reliable and efficient offline analysis of this class of time series. The inference is carried out through a three-step approach: detecting the structural changes of the time series using a recently proposed multiwindow algorithm, identifying each segment as a state and selecting the most appropriate number of states, and estimating the Markov source based upon the symbolic sequence obtained from previous steps. We provide theoretical results and algorithms in order to facilitate the inference procedure described above. We demonstrate the accuracy, efficiency, and wide applicability of the proposed algorithms via an array of experiments using synthetic and real-world data.**

***Index Terms*—Consistency, multi-regime models, prediction, recurring patterns, time series.**

I. Introduction

MODELING and forecasting time series is of fundamental importance in a variety of applications. Temporal measurements collected from various domains usually exhibit occasional changes and recurring patterns, and therefore they cannot be modeled as a single stationary process. Some examples are monthly temperature, hourly electricity demand in a city, U.S. business cycles [1], Canadian lynx series [2], electroencephalogram (EEG) signals [3], environmental measurement [4], and activity data collected from wearable devices [5].

A recurring pattern offers non-negligible predictive power. Suppose, for example, that an EEG recording exhibits a particular pattern leading up to the onset of an epileptic seizure. Then one might reasonably expect that it is possible to predict the next ictal event by modeling historical data. Long-period cyclicality cannot be well modeled by a single time-series model such as autoregression. It is natural to consider a model that occasionally varies over time and that can recur. To address this, we assume that there exists a dictionary of models, and that one such model best explains the data for a given time epoch.

In line with this idea, a generic model is to assume that the observation x_t at each time step t is a random variable with probability density function (PDF) $p_t(x_t \mid x_{t-1}, x_{t-2}, \ldots)$, where p_t comes from a finite set of PDF's, also referred to as states. When the states are assumed to follow a first order Markov process, the time series is usually referred to as a regime-switching model [6], [7]. Note that if each row of the transition probability matrix is the same, the states then follow independent multinomial distributions. Therefore, it is common to assume the Markov chain to be quite persistent [8]. If the data within each segment is further assumed to be generated from a parametric family, then a fully parametric model can be established. In fact, the case where x_t takes values from a fixed set of constants is usually referred to as a Hidden Markov model, and its analysis can be traced back to [9], [10]. The model with general p_t's has been studied in the speech recognition literature [11]. A similar idea also led to self-exciting threshold autoregressive models [12], where the states are triggered by historical values of x_t. When the states are assumed to be periodic, time-series models such as periodic autoregression are commonly utilized [13], [14]. A class of nonparametric piecewise autoregressive models was studied in [15], where the inference procedure was cast as an optimization problem. Other attempts to characterize these types of models follow optimization perspectives [16]–[18], the model averaging perspective [19], or Bayesian perspectives under which the number of states and parameters are jointly inferred using Markov chain Monte Carlo techniques [20]–[22]. A comprehensive review on time-varying processes can be found in [23]. The aforementioned methods have been applied to several interesting applications such as forensics [24], transportation [25], energy [26], [27], neuroscience [28], [29], and finance [30]–[35].

As with other statistical inference tasks, any prescribed parametric model here may not be adequate to explain the observed data, as the model may be subject to aperiodic changes in behavior. To overcome this shortcoming, a suitable approach is to postulate multiple parametric models and apply a model selection procedure to select the optimal one. In this paper, we aim to address this issue as follows. Given a collection of parametric probability distribution functions p_t, we provide a scheme that accurately determines the number of regimes, denoted by s. The key challenge is that models with larger s fit the observed data better. However, using larger s has the unfortunate consequence of over-fitting, which decreases the predictive power. On the other hand, smaller s might give rise to

Manuscript received May 29, 2017; revised October 14, 2017 and December 1, 2017; accepted February 8, 2018. Date of publication March 12, 2018; date of current version April 2, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Huang. This work was supported by the Defense Advanced Research Projects Agency under Grant W911NF-14-1-0508 and Grant N66001-15-C-4028. *(Corresponding author: Jie Ding.)*

J. Ding and V. Tarokh are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: djrlthu@gmail.com; vahid.tarokh@duke.edu).

S. Shahrampour and K. Heal are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: shahin@seas.harvard.edu; kathrynheal@g.harvard.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2018.2811757

a insufficient description of the model which is also unable to predict accurately. Therefore, an appropriate selection of s is essential.

A relevant framework for detecting the optimal number of states is to apply likelihood ratio tests, i.e., to test the null hypothesis that there are s states, versus the alternative hypothesis that s is larger, for $s = 1, 2, \ldots$. However, this approach is not theoretically founded because of the lack of regularity conditions [36] for time series with mixture-model structures. Another approach is to apply a maximum penalized likelihood method such as the Akaike Information Criterion (AIC) [37], [38] or Bayesian Information Criterion (BIC) [39]. As an example, recent works have extended the applicability of AIC and BIC to finite mixture models, including the evaluation of AIC for mixture regression models [40], the introduction of a modified AIC for mixture regression models [41], and BIC-like consistent criterion for mixture models [42]. A comprehensive review on information criteria can be found in [43]. However, the existing theories have been established for a single time-series model, and it is not clear whether they can be extended to mixture models in the presence of data-dependent structure.

In order to select the number of states s, our first contribution is to propose a test statistic that is asymptotically chi-square distributed under the null hypothesis. This newly-proposed statistic allows for identification of the true number of states via hypothesis testing. An analytical difficulty stems from the fact that sample points (data) across various states are mutually dependent. Our next contribution is to find theoretical guarantees (proper penalty function) under which the k-means algorithm is capable of simultaneously identifying the state of each segment correctly and consistently.

In practice, we are often interested in persistent states [8]. A commonly used model is a Markov switching model with transition probabilities close to zero. A characteristic of this model is that its true parameters are located near the boundary of the parameter space. Thus when using this model, one may encounter errors due to a failure of the "interior point" regularity assumption, commonly used to prove asymptotic results (see for example [42], [44]). What is more, enumerating candidate models also requires large computational cost. Our novel approach overcomes these challenges; it is conceptually intuitive and computationally tractable. The idea is to detect the switching point first, and then fit a high-order Markov model by adding some sparsity constraint. The sparsity is imposed through an information-theoretic coding method that alleviates the aforementioned "interior point" issue. Our method also serves as a preliminary step to propose more suitable parametric models, as will be discussed more in the experiment section.

The remainder of the paper is outlined below. In Section II, we introduce the background and model. After that, we propose a three-step approach to analyze multi-state time series in Section III, and propose our statistical test to identify the number of states, followed by our theoretical results. In Section IV, we evaluate the proposed modeling technique with experiments. We make our conclusions in Section V, and include the proofs of main results in the Appendix.

II. Multi-state Autoregressive Models

We first introduce some notation for the rest of the paper. Let $\mathcal{N}(\mu, V), \chi^2_k, \mathcal{M}(\boldsymbol{p})$, respectively denote the normal distribution with mean μ and covariance matrix V, the chi-square distribution with k degrees of freedom, and the multinomial distribution with parameter $\boldsymbol{p}$. This $\mathcal{M}(p)$ is represented by a nonnegative vector with unit 1-norm. Given an $s \times s$ positive definite matrix V, let $\|\cdot\|_V$ denote the norm $\|y\|_V^2 = y^{\mathrm{T}} V y$ for all $y \in \mathbb{R}^{s\times 1}$. We use $\rightarrow_p$ and $\rightarrow_d$ to denote the in probability and in distribution convergence.

A sequence of data $\{x_t, t = 1, 2, \ldots\}$ is generated by an autoregressive model of order $L \in \mathbb{N}$, also referred to as AR(L), if

$$x_t = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_t + \varepsilon_t,$$

where $\boldsymbol{x}_t \triangleq [x_{t-1}, \ldots, x_{t-L}]^{\mathrm{T}}$, $\boldsymbol{\beta} \triangleq [\beta_1, \ldots, \beta_L]^{\mathrm{T}}$, and ε_t are independent and identically distributed (i.i.d.) random variables with zero mean and variance σ^2 [45]–[47]. We sometimes refer to the vector $\boldsymbol{\beta}$ as an AR filter of order L. In particular, AR(0) denotes a sequence of i.i.d. noise variables, namely $x_t = \beta_0 + \varepsilon_t$. Note that we use $x_{1-\ell}, \ell = 1, \ldots, L$ to denote initial observations, where subscript t represents the data at time step t.

In this work, we restrict our attention to the space of all stable AR filters, denoted by R_L. By "stable" we mean that if $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_L]^{\mathrm{T}} \in R_L$, then all of the (complex) roots of its characteristic polynomial $z^L - \sum_{\ell=1}^{L} \beta_\ell z^{L-\ell}$ lie inside of the unit circle.

What has been presented so far describes the class of *single-state* AR models, where the coefficient $\boldsymbol{\beta}$ remains constant throughout the process. In this work, however, we are interested in analyzing *multi-state* AR models, characterized by a time-varying coefficient β_t, as follows:

$$x_t = \beta_{t,0} + \sum_{\ell=1}^{L} \beta_{t,\ell} x_{t-\ell} + \varepsilon_t.$$

Here the AR filter $\boldsymbol{\beta}_t \triangleq [\beta_{t,0}, \ldots, \beta_{t,L}]^{\mathrm{T}}$ at each time step t belongs to a finite set $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{s_0}\}$, and ε_t are independent noise variables. If we let $\boldsymbol{\beta}_t = \boldsymbol{\gamma}_{z_t}$, then $z_t \in \{1, \ldots, s_0\}$ is called the state at time t. Two underlying assumptions in this model are: (1) Each data point depends linearly on finitely many previous points, corrupted by a random noise. The assumption of dependence only on previous points is reasonable since the observations are obtained sequentially in time. (2) There are finitely many AR states, which is reasonable if the stochastic process exhibits different patterns in different epochs, and each pattern recurs. For example, stock prices may fall into a few regular patterns throughout business cycles. The multi-state AR model serves as a generic model that can be used to fit data in various real-world applications.

A purely parametric model requires further assumptions on the data generating process of (hidden) variables z_t. For example, when one assumes that the states are generated independently and according to a multinomial distribution, the model is capable of representing non-linear and non-stationary time series with multimodal conditional distributions and with

heteroscedasticity [48]. As another example, one may assume that $\{z_1, z_2, \ldots\}$ follows a Markov chain, or that the switches of states follow a point process parameterized by exogenous variables. In those cases, the maximum-likelihood estimator (MLE) for the mixture model may be obtained via, for instance, the expectation-maximization algorithms. However, the usual assumption that states follow a first order Markov process may be too restrictive. Instead of postulating a fully parametric model, we allow the states to follow a Markov process of unknown high order.

We propose the use of a three-step strategy that infers the parameters in each state and the transition of states separately. Our basic idea is to divide the inference procedure into three steps: 1) discover the switch points using a change detection approach and estimate the parameters within each segment; 2) identify the segments associated with the same state since each state can be revisited; and 3) estimate how the states transition.

While we briefly discuss all three steps in this paper, our focal point is the second step, where we provide identification algorithms supported with theoretical and empirical guarantees. The first step can be tackled using change point detection algorithms; we developed one such algorithm in a companion paper [49]. Furthermore, we introduce the Context Tree Weighting (CTW) method for the third step.

III. Methodology & Results

A. Step 1: Identifying the Change of States

A typical offline multiple change point analysis aims to solve the following problem. Given observations $y_1, \ldots, y_T \in \mathbb{R}^d$ and some $m \in \mathbb{N}$, the goal is to find integers $0 < \ell_1 < \cdots < \ell_m < T$ that minimize the sum of within-segment loss

$$e_m \triangleq \sum_{k=1}^{m+1} Loss(y_{\ell_{k-1}+1}, \ldots, y_{\ell_k}), \tag{1}$$

where $Loss(\cdot)$ is an appropriate loss function, $\ell_0 = 0$, and $\ell_{m+1} = T$. Here, the number of change points m is usually determined by a two-step penalized approach: first minimize (1) for different m's, and then select the optimal one by minimizing $e_m + mf_T$, where f_T is a proper penalty term. A simple and widely adopted loss function for independent data (with different means in each segment) is the quadratic loss

$$Loss_q(y_{\ell_{k-1}+1}, \ldots, y_{\ell_k}) \triangleq \sum_{t=\ell_{k-1}+1}^{\ell_k} \|y_t - \bar{y}_k\|_2^2, \tag{2}$$

where $\bar{y}_k$ is the sample mean of $y_{\ell_{k-1}+1}, \ldots, y_{\ell_k}$.

For general loss functions, the bisection procedure [50], [51] and exact search methods such as segment neighborhood [52], optimal partitioning [53], [54], and the PELT method [55] have also been widely applied. For the special case where a time series is segmented into several parts, each modeled by an autoregressive model, a multi-window (MW) algorithm was proposed in [49] that achieves both accurate change detection and near linear computational cost, outperforming the state-of-the-art bisection procedure. More details of this algorithm, as well as relevant MATLAB software, can be found in [49]. We will provide here a brief overview of the MW algorithm.

Assuming that the order L is known *a priori*, the multi-window approach uses a sequence of R window sizes $w_1 > \cdots > w_R$ in order to capture true segments regardless of size. For each w_r, the algorithm

1) transforms the original data into a sequence of $L+1$ dimensional points, all of which are asymptotically independent;
2) minimizes (1) based on the quadratic loss and applies penalized model selection to determine the optimal number of change points;
3) maps the change points output from step 2 back to the original scale $\{1, \ldots, T\}$ and obtains several short ranges (intervals) $I_k^{(r)}, k = 1, 2, \ldots$ (each of size $2w_r$) that are likely to contain the desired change points.

After repeating the above procedure for different window sizes w_r, the MW algorithm combines the information using the following rule: the detected intervals of change points from each window size are each assigned a score of one, those scores are aggregated, and those highest-scoring intervals are selected as change points. An optional post-processing step can be applied to find the exact optimal change points. We refer to [49] for experiments on various datasets and theoretical results. It has been shown in [49] that the identified number of change points converges almost surely to the truth (if it exists) under reasonable assumptions, and the deviation between the detected change points is negligible as data size tends to infinity.

B. Step 2: Identifying the States

In view of Section III-A, in this section we assume that a set of potential change points has been already discovered. In practice, Step 1 serves as a preliminary screening procedure to facilitate the discovery of change points and then the labeling of states. It remains to identify which segments are generated by the same AR model. This section consists of two parts. First, we propose a test statistic that is asymptotically chi-square distributed under the null hypothesis, that is, two segments are generated from the same state. Second, we prove that a direct application of the *k*-means algorithm is capable of simultaneously identifying the state of each segment correctly and consistently. Note that our theory will be applicable in the presence of a superfluous change point. That is, state identification will succeed even if one begins Step 2 with the assumption that two neighboring segments, truly generated from the same state, are distinct.

We write the matrix representation of the single-state AR model as $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\boldsymbol{y} \triangleq [x_1, \ldots, x_T]^{\mathrm{T}}$$
$$X^{\mathrm{T}} \triangleq [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$$
$$\boldsymbol{\varepsilon} \triangleq [\varepsilon_1, \ldots, \varepsilon_T]^{\mathrm{T}}.$$

The least squares estimation of $\boldsymbol{\beta}$ from $x_1, \ldots x_T$ is

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y},$$

and the sum of squared error (SSE) is

$$\begin{aligned}\hat{e} &= \|y - X\hat{\boldsymbol{\beta}}\|_2^2 \\ &= \|(I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})\boldsymbol{y}\|_2^2 \\ &= \|\varepsilon\|_{I-P_X}^2 ,\end{aligned}$$

where $P_X \triangleq X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ denotes the projection matrix associated with X. We may also write $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}}_{1:T}$, where the subscript $1:T$ is used to emphasize the dependency on samples $x_1, \ldots, x_T$ (in addition to L implicit initial values). When the constant term is taken into account, the least squares estimates are similar except that X is replaced with $\underline{X}$, where $\underline{X} \triangleq [\mathbf{1}, X]$ with $\mathbf{1}$ denoting the column vector of all ones.

Suppose that data consists of $m+1$ ($m \geq 1$) segments, denoted by $\{x_{T_{k-1}+1}, \ldots, x_{T_k}\}_{k=1}^{m+1}$, where $0 = T_0 < T_1 < \cdots < T_{m+1} = T$. Least squares estimates are applied to each segment, with the SSE denoted by $\hat{e}_{T_{k-1}+1:T_k}$, $k = 1, \ldots, m+1$. Since the fitting error of a model is no larger than that of a smaller model that nests in the larger one, it is clear that $\hat{e}_{1:T} \geq \sum_{k=1}^{m+1} \hat{e}_{T_{k-1}+1:T_k}$. We define the overfitting gain by

$$\eta_m \triangleq \log \frac{\hat{e}_{1:T}}{\sum_{k=1}^{m+1} \hat{e}_{T_{k-1}+1:T_k}}, \tag{3}$$

and a normalized gain

$$g_m = T\eta_m .$$

By its definition, the overfitting gain quantifies the advantage of modeling data by $m+1$ segments over one segment. We note that both g_m and η_m are invariant to a scaling of noise variance σ^2. We provide the following result characterizing the asymptotic behavior of the normalized gain.

Theorem 1: Suppose that $x_1, \ldots, x_T$ are generated from a zero mean stable AR(L) model, and noise variables ε_t are i.i.d. with zero mean, variance σ^2, and finite fourth moments. Assume that $T_k - T_{k-1} \to \infty$ as $T \to \infty$, $k = 1, \ldots, m+1$ ($m \geq 1$). Suppose also that the distribution of ε_1 has a nontrivial absolutely continuous component, and that $E[\max\{(\log|\varepsilon_1|), 0\}] < \infty$. Then $g_m \to_d \chi^2_{mL}$ as $T \to \infty$.

Before providing insights into Theorem 1, we state two corollaries of that result, which will require us to first define

$$g'_m \triangleq \hat{e}_{1:T} - \sum_{k=1}^{m+1} \hat{e}_{T_{k-1}+1:T_k} .$$

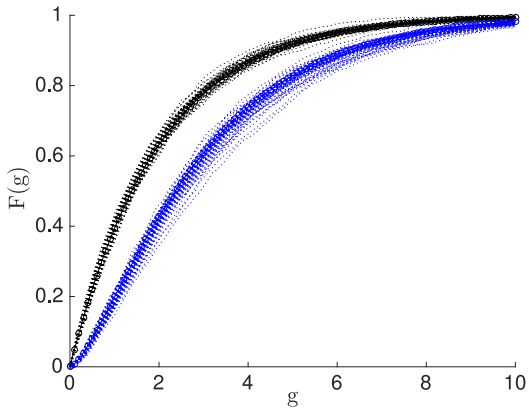
A byproduct of the proof of Theorem 1 implies the following corollary.

Corollary 1: Under the same conditions of Theorem 1, $g'_m/\sigma^2 \to_d \chi^2_{mL}$ as $T \to \infty$.

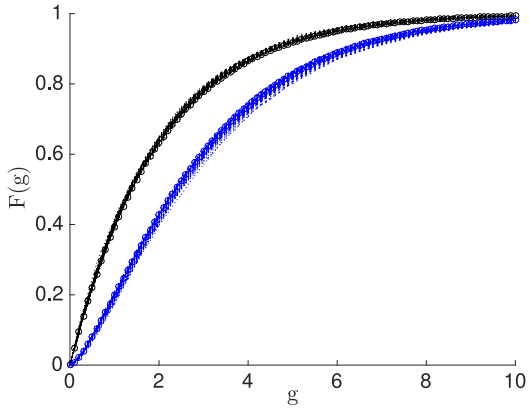
We could take into account the constant term in applying the least squares, namely to use $\underline{X}$ instead of X as the design matrix. In that case, the overfitting gains are denoted by $\underline{g}_m, \underline{g}'_m$. Following similar proof of Theorem 1, we have the following corollary.

Corollary 2: Under the same conditions of Theorem 1, $\underline{g}_m \to_d \chi^2_{m(L+1)}$, and $\underline{g}'_m/\sigma^2 \to_d \chi^2_{m(L+1)}$ as $T \to \infty$.

Remark 1: The assumption on noise is mild. For instance, it is satisfied by distributions such as Gaussian, doubled exponential (Laplace), doubled chi-square, etc. Remarkably, Theorem 1 implies that the asymptotic distribution does not depend on the locations of splitting point T_k and the true filter $\boldsymbol{\beta}$. In addition, synthetic data experiments show that the empirical distribution is close to the asymptotic distribution even for finite sample size (small T). To illustrate, we draw 50 AR(2) filters uniformly from R_L (using Algorithm 3 in [56]); for each AR(2), we generate an AR sequence of $T = 150$ data and record the overfitting gain g_1 with $T_1 = 50$, and repeat a few independent random instances to obtain a empirical CDF of g_1 (plotted in black dashes of Fig. 1). This empirical CDF is close to the CDF of χ^2_2 plotted in black dotted lines. We also repeat the experiments for $\underline{g}_1$ (in blue).



(a) 10^3 independent draws



(b) 10^4 independent draws

Fig. 1. Figures showing the empirical CDF of g_1 (in black dotted lines) and $\underline{g}_1$ (in blue dotted lines) computed from several independent draws with $T_1 = 50, T = 150$, for each of the fifty AR(2) filters uniformly generated from R_L, along with the true CDF of χ^2_2 (in a black circled line) and of χ^2_3 (in a blue circled line).

The result can be used for a hypothesis test for whether two (or more) segments are associated with the same state. For example, suppose that there are two segments $x_{T_{k-1}+1:T_k}$ and $x_{T_{\ell-1}+1:T_\ell}$ ($1 \leq k < \ell \leq m+1$), associated with state labels z and z', respectively. Consider the hypothesis test

$$H_0 : z = z', \quad H_1 : z \neq z'. \tag{4}$$

Algorithm 1: State Identification.

input $\hat{\beta}_k, k = 1, \ldots, m+1$, penalty f_T.
output $\hat{s}, \hat{z}_k, k = 1, \ldots, m+1$
1: **for** $s = 1 \to m+1$ **do**
2: Apply the *k*-means algorithm to obtain s clusters from $\hat{\beta}_k, k = 1, \ldots, m+1$, with the sum of within-cluster squared distance denoted by ℓ_s.
3: **end for**
4: Let $\hat{s} = \arg\min_{1 \le s \le m+1}\{\ell_s + s f_T\}$, and $\hat{z}_k \in \{1, \ldots, \hat{s}\}, k = 1, \ldots, m+1$ be the associated labels.

We compute

$$g_1 \triangleq (T_k - T_{k-1} + T_\ell - T_{\ell-1}) \log\{\hat{e}/(\hat{e}_{T_{k-1}+1:T_k} + \hat{e}_{T_{\ell-1}+1:T_\ell})\},$$

where $\hat{e}$ denotes the SSE if the two segments are merged. Given a significance level $\alpha \in (0,1)$, we compare g_1 with $t_L \triangleq F^{-1}_{\chi^2_L}(1-\alpha)$ where $F^{-1}_{\chi^2_L}(\cdot)$ denotes the inverse function of the CDF of χ^2_L. We reject H_0 if $g_1 > t_L$. If we take into account a possibly nonzero mean, we use $\underline{g}_1$ and t_{L+1} instead.

Remark 2: Using Theorem 1 or its corollaries, we can apply the above hypothesis test to each pair of segments to determine whether they are from the same state. However, this procedure can be computationally cumbersome if there are a large number of segments. To alleviate the issue, we can apply the *k*-means clustering algorithm to the estimated AR filters to simultaneously label their states, instead of pairwise tests (Algorithm 1). We show in the result below that a proper penalized approach produces correct labelling of states. Illustrative experiments are given in Section IV, where we let f_T be $L\log(\tilde{T})/\tilde{T}$ (which resembles BIC) is used by default, with $\tilde{T}$ defined in the statement of Theorem 2. Note that the theorem applies even if two neighboring segments are truly from the same state.

Before we proceed, we briefly discuss the issue of identifiability of states. In the multi-state model, the parameters (AR coefficients) are identifiable only up to a permutation of states. We need a slightly extended notation of identifiability in the sense of an equivalence relation. We will say that $\{z_1, \ldots, z_T \in \mathcal{A}\} \equiv \{z'_1, \ldots, z'_T \in \mathcal{A}'\}$ if and only if $\mathcal{A} = \mathcal{A}'$ and $z_t = \pi(z'_t)$ $(t = 1, \ldots, T)$ for some permutation $\pi : \mathcal{A} \mapsto \mathcal{A}$.

Theorem 2: Suppose that $m+1 \ge s_0$ and that the assumptions of Theorem 1 hold. Let $\tilde{T} \triangleq \min_{1 \le k \le m+1}(T_k - T_{k-1})$. Then for any f_T and $\tilde{T}$ satisfying

$$\lim_{T\to\infty} f_T + \frac{1}{\tilde{T} f_T} = 0, \tag{5}$$

The selected number of states $\hat{s}$ in Algorithm 1 satisfies $\hat{s} \to_p s_0$ as $T \to \infty$.

While in Algorithm 1, our clustering is based on the AR coefficients, we remark that an alternative approach is to cluster based on the roots of the characteristic polynomial which determine the physical behavior of AR systems. In other words, there exist two perspectives for measuring the distance of time-series models: time domain and frequency domain. Our focus in this work is on the former.

C. Step 3: Modeling the State Transitions

The Context Tree Weighting (CTW) method is a sequential universal data compression procedure for a sequence with finite alphabet, which works by mixing the predictions of many underlying variable order Markov models [57], [58]. We propose to use it to model and predict the state sequence $\{z_1, \ldots, z_t\}$ (also denoted by $z_{1:t}$), where $z_t \in \mathcal{A}$, and $\mathcal{A} = \{1, \ldots, s\}$ is the finite alphabet (labels of states).

Following standard notation from information theory, we sometimes refer to a sequence $\{z_1, \ldots, z_k\} \in \mathcal{A}^k$ as a string, also denoted by $\boldsymbol{z} \triangleq z_1 \cdots z_k$. We say $\boldsymbol{z}$ is a suffix of another $\boldsymbol{z}' \triangleq z'_1 \cdots z'_{k'}$ if $k \le k'$ and $z_{k-i} = z'_{k'-i}, i = 0, \ldots, k-1$. The empty string, denoted by $\emptyset$, is the suffix of all strings. A proper and complete suffix set, denoted by $\mathcal{V}$ with memory no larger than $d \in \mathbb{N}$ is defined below. This is a set of strings in the form of $[z_1, \ldots, z_k]$ where $z_k \in \mathcal{A}, k \le d$, which satisfies the following conditions: 1) no string in $\mathcal{V}$ is a suffix of any other string in $\mathcal{V}$; and 2) for any $t \in \mathbb{N}$, any string $z_1 \cdots z_t \in \mathcal{A}^t$ has a suffix that belongs to $\mathcal{V}$ (this suffix is unique since $\mathcal{V}$ is proper). Following the definition, any proper and complete suffix set induces a function $\mathcal{F} : z_1 \cdots z_t \mapsto v \in \mathcal{V}$. Let $p(v)$ denote a discrete measure over $\mathcal{A}$ for each $v \in \mathcal{V}$. We assume the following true data generating process.

The sequence $z_{1:t}$ is generated by a finite memory tree source defined by $z_t \mid z_{(t-d):(t-1)} \sim \mathcal{M}(\boldsymbol{p}_t)$, where $\boldsymbol{p}_t \triangleq \boldsymbol{p}(\mathcal{F}(z_{t-d:t-1}))$, $z_{1-d}, \ldots, z_0$ are known initial values.

Let $D_{KL}(f, g)$ denote the Kullback-Leibler divergence of PDF $f(\cdot)$ from $g(\cdot)$. Because the CTW exploits the potentially sparse tree structure, it can be much more efficient than maximum likelihood estimation when directly applied to a Markov process, especially when the Markov order is large. Given d initial states, the CTW procedure sequentially produces an estimate of the joint distribution of $z_{1:t}$, $P_w(z_{1:t})$, which is shown to converge to the stationary probability measure $P_a(z_{1:t})$ in the sense that $D_{KL}(P_a(z_{1:t}), P_w(z_{1:t}))/t \to 0$ as $t \to \infty$ [58].

By using this strategy, we do not assume specifically how the states evolve, but we need to assume that each state is fairly persistent, i.e., the time intervals between two consecutive state changes are long. This is a reasonable assumption in many real cases such as EEG data, speech data, and environmental data where the sampling rate is high.

IV. Numerical Experiment

A. Synthetic Data Experiment: Three-step Approach

We illustrate the three-step approach via a numerical experiment. We define 3 states (labeled $\mathcal{A} = \{1, 2, 3\}$) as zero mean AR(2) $x_t = \boldsymbol{x}_t^{\mathrm{T}} \boldsymbol{\beta}^{(k)} + \varepsilon_t$, $k = 1, 2, 3$, with independent $\mathcal{N}(0,1)$ noises, $\boldsymbol{\beta}^{(1)} = [0.8, -0.5], \boldsymbol{\beta}^{(2)} = [-0.6, -0.7]$, and $\boldsymbol{\beta}^{(3)} = [0, 0.6]$. We generate a time series of length $T = 3000$ that consists of 20 segments $x_{T_{k-1}+1:T_k}$, $k = 1, \ldots, 20$, corresponding to the state sequence $1, 2, 3, 2, \ldots, 1, 2, 3, 2$ (of length 20). The segment lengths $T_k - T_{k-1}, k = 1, \ldots, 20$ are generated from the symmetric Dirichlet distribution with concentration parameter 10. The above procedure is repeated

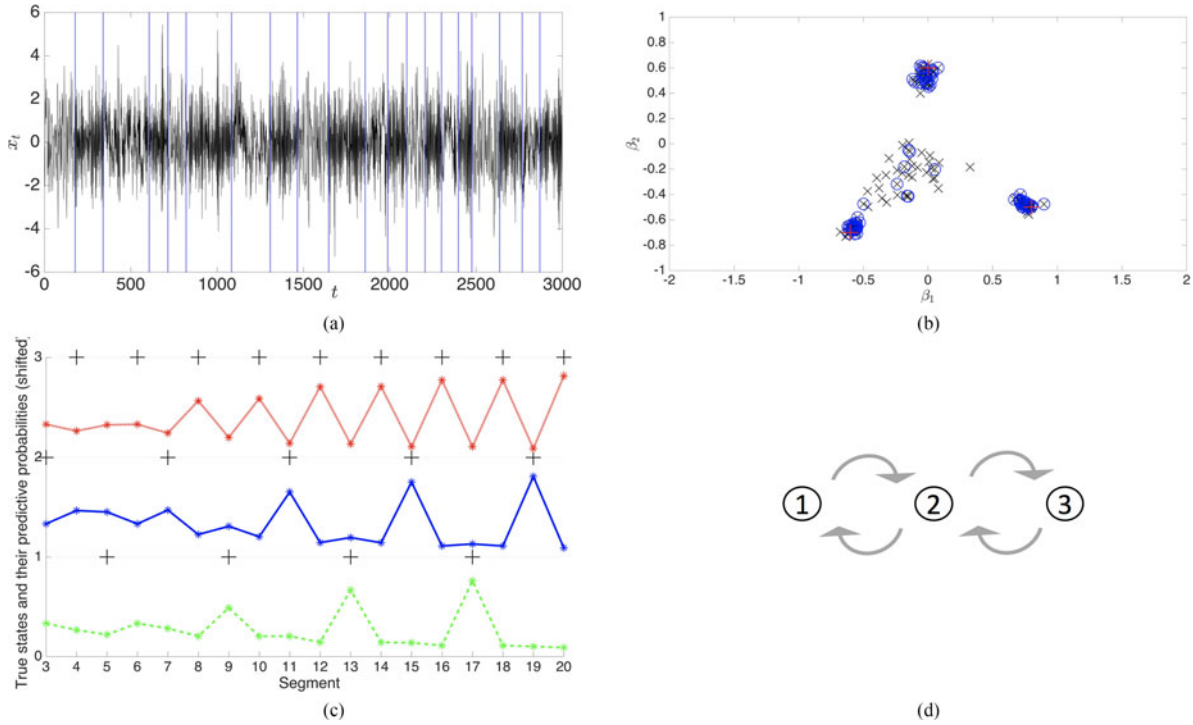


Fig. 2. A graph showing (a) one of the 50 time series (black line) with the true locations of state transitions (vertical blue dash); (b) the estimated coefficients $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2$ of each experiment (black cross), together with those coming from the 29 experiments with correctly identified number of states (blue circle), and with the true filters (red plus); (c) the predictive probabilities of observing state $i \in \{1, 2, 3\}$ plus $i - 1$ (in green dash, blue line, red small dash, respectively), along with the true states (black plus); (d) the recovered patten of state transitions.

50 times. In each repetition, we apply the three-step approach in Section III and summarize the final results in Fig. 2.

Fig. 2(a) depicts one instance of the data (in black line) with the true locations of state transitions (in blue dash). We first apply the MW algorithm to obtain the change points. We then apply Algorithm 1 (Section III-B, with the default choice of $f_T = 2\tilde{T}^{-1}\log(\tilde{T})$, $\tilde{T}$ being the minimal segment length from the last step) to obtain the optimal number of states and their labels. The estimated coefficients $\hat{\beta}_1, \hat{\beta}_2$ of 50 experiments are plotted together in Fig. 2(b). The coefficients that are associated with the experiments where the number of states is correctly identified (29 in total) are highlighted with circles. In the plot, the cloud of estimated filters are centered around the true filters. We notice a slight "pooling" of estimates toward each other and the presence of a few outliers. The pooling effect occurs due to minor but unavoidable error in the estimator, i.e., when the discovered change points do not coincide precisely with true state transitions. The outliers result from false negatives, i.e. when the estimator misses a true change point (which causes a reconciliation between adjacent filters). The latter issue can be alleviated by deliberately overfitting the number of change points (in implementing Section III-A). Lastly, we apply the CTW approach with depth $d = 2$ to sequentially predict the next state. We define the state in segment level in the sense that if $m - 1$ switching points are discovered, the state sequence is of m.

As was discussed in Remark 2, the states are identifiable only up to permutations. For the sake of easy presentation, we relabel the states in the following way. For those experiments that correctly identify the number of states, we find the permutation $\pi : \mathcal{A} \mapsto \mathcal{A}$ such that $\sum_{i\in\mathcal{A}} \|\boldsymbol{\beta}^{(\pi(i))} - \hat{\boldsymbol{\beta}}^{(i)}\|_2^2$ is minimized, and relabel i as $\pi(i), \forall i \in \mathcal{A}$. After relabeling, the identified state sequences (in segment level) in all of the 29 experiments are the correct sequence (namely 1, 2, 3, 2, ...). The predictive probabilities of observing state $i \in \{1, 2, 3\}$ plus $i - 1$ are plotted in Fig. 2(c) along with the true states. The closer the probability (curve) is to the true state (point), the better predictive performance we achieved. The pattern of state transitions is soon accurately characterized, as illustrated in Fig. 2(d).

To verify that the performance of Algorithm 1 improves as sample size increases, we repeat the experiment of the three-step approach for $T = 10^3, 3 \times 10^3, 5 \times 10^3, 10^4$, where the average selected number of states (with standard errors) are respectively $2.26(0.13), 3.62(0.15), 3.22(0.09), 3.10(0.10)$. To further focus on the verification of Theorem 2 (proved in the Appendix), we simplify the setting by first assuming that the locations of change points are known, and then applying Algorithm 1 for the same synthetic data with $T = 100, 300, 500, 1000, 3000, 10000$, each with 50 independent repetitions. In addition to the average selected number of states $\hat{s}$, we also recorded the measures

$$\hat{u} \triangleq \frac{\text{card}\{(i,j) : a_i \neq a_j, \hat{a}_i = \hat{a}_j, 1 \leq i, j \leq 20\}}{\text{card}\{(i,j) : a_i \neq a_j, 1 \leq i, j \leq 20\}},$$
$$\hat{o} \triangleq \frac{\text{card}\{(i,j) : a_i = a_j, \hat{a}_i \neq \hat{a}_j, 1 \leq i, j \leq 20\}}{\text{card}\{(i,j) : a_i = a_j, 1 \leq i, j \leq 20, i \neq j\}}, \quad (6)$$

TABLE I
PERFORMANCE OF ALGORITHM 1

T	100	300	500	1000	3000	10000
$\hat{s}$	8.40	4.94	3.92	3.10	3.00	3.00
	(0.181)	(0.144)	(0.137)	(0.055)	(0.000)	(0.000)
$\hat{u}$	0.06	0.01	0.001	0.0002	0.00	0.00
	(0.005)	(0.002)	(0.0005)	(0.0002)	(0.000)	(0.000)
$\hat{o}$	0.76	0.36	0.16	0.02	0.00	0.00
	(0.011)	(0.025)	(0.024)	(0.011)	(0.000)	(0.000)

along with their standard errors in parentheses, where $\{a_i\}_{i=1}^{20}, \{\hat{a}\}_{i=1}^{20}$ are the true and the estimated state sequences. Larger $\hat{u}$ (resp. $\hat{o}$) implies more underfitting (resp. overfitting). In this simulation, we use Laplace noises, and give the results in Table I. As shown in Table I, the number of states and labeling are correctly identified as T grows large.

B. Synthetic Data Experiment: Comparison with a Fully Parametric Model

The purpose of this section is to compare our three-step approach for offline observations to the maximum likelihood estimation commonly used for statistical inference of parametric models. We show that our approach can serve as a fast preprocessing method for classical inference procedures. Suppose that the time-series model consists of 3 states, corresponding to zero mean AR(2) with independent $\mathcal{N}(0,1)$ noises, and coefficients $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}$ i.i.d. samples from the uniform distribution on the space of all stable AR(2) filters (using Algorithm 3 introduced in [56]) for each run of the experiment. The law of state transitions follows a Markov process, with transition probability matrix T being $T_{ii} = 1 - c, T_{ii'} = c/2, c = 0.01,\ \forall i, i' = 1, \ldots, 3, i \neq i'$. In other words, we may write the data generating process as

$$z_t \sim \begin{cases} \mathcal{M}(\pi_1, \ldots, \pi_M) & \text{if } t = 1, \\ \mathcal{M}(T_{j1}, \ldots, T_{jM}),\ j = z_{t-1} & \text{otherwise}, \end{cases} \tag{7}$$

$$x_t \sim \mathcal{N}(\boldsymbol{x}_t^{\mathrm{T}} \boldsymbol{\beta}^{(z_t)}, \sigma^2_{z^{(t)}}), \quad t = 2, \ldots, T. \tag{8}$$

where $z_1, \ldots, z_T$ denote the sequence of unobserved states.

We generate synthetic data from the above model and run the three methods described next. The first method considered is the three stage approach (denoted by "3-step"). The estimated transition probability from state i' to i is taken to be the latest CTW predictive probability after observing i' $(1 \leq i', i \leq 3)$. We also use the standard expectation-maximization (EM) algorithm to numerically find the maximum likelihood estimator (MLE). The derivation of the EM algorithm for the above particular model can be found in [56, Sec. 3]. We assume that the number of states is known a priori for the EM algorithm. The initial values of unknown parameters are chosen to be $\pi_i = 1/3, \sigma_i^2 = 10, \boldsymbol{\beta}^{(i)} \sim \mathcal{N}(0, I), i = 1, 2, 3$. The drawback of the EM algorithm is that its convergence speed and performance strongly depend on its initialization. An improper initialization causes EM to converge to a local maximum that can be quite far from the global optimum (which is the MLE). A popular technique is to use multiple random initializations and choose the output with the largest likelihood [59], but this can be quite time consuming. Instead, we first apply "3-step" and use the obtained results as the initial values for EM in order to improve the convergence. We refer to the hybrid approach as "3-step-EM".

We repeat the experiment 50 times independently, for each $T = 10^3, 3 \times 10^3, 10^4$. The results are summarized in Table II. We compare the mean square error (MSE) of the estimated coefficients (average $\sum_{i=1}^{3} \|\boldsymbol{\beta}^{(i)} - \hat{\boldsymbol{\beta}}^{(i)}\|_2^2$), MSE of the transition probabilities (average $\sum_{i',i=1}^{3} (T_{i',i} - \hat{T}_{i',i})^2$), and the computational cost (in seconds). Each mean value is provided with its standard error in the parentheses below. The algorithms were implemented in MATLAB and run on a PC with 3.1 GHz dual-core CPU. To overcome the identifiability issue, a relabeling has been applied that is similar to the last section.

As expected, EM with calibrated initialization ("3-step-EM") gives the lowest MSE of coefficients. In addition, "3-step" performs better than other two approaches in terms of the estimation of transition probabilities and computational cost, and slightly better than EM (with random initialization) in terms of estimating AR coefficients.

Fig. 3 shows one instance of the time series, along with the true states and the discovered states by three approaches. The estimated state $\hat{z}_t$ at time step t for the latter two approaches is defined as the state with the largest posterior probability conditioned on the observations. Fig. 3 demonstrates the typical performance of three methods. We have noticed that "3-step" tends to produce overly persistent states. This phenomenon can be avoided by either fine tuning the penalization used to determine the number of change points in MW algorithm (we had been using the default option suggested in [49]), or using an exact search algorithm (more time consuming) such as the PELT method [55]. On the contrary, "EM" tends to capture the true transitions, but it tends to produce states with overly frequent transitions. This occurs because the posterior means of the hidden states exhibit variations and EM does not encourage them to be persistent. Finally, "3-step-EM" strikes a satisfying balance between the previous two methods.

In summary, our proposed approach is comparable to the classical strategies used for inferring a particular mixture model. In addition, it can serve as a companion of classical inference procedures to further enhance their performance. Finally, our approach does not require any assumption on parametric forms, so it is generally applicable.

C. Real Data Experiments

1. Eastern US temperature data: In this experiment, we revisit the temporal variability of the summer temperatures over the Eastern US for 1895-2015, which was also studied in [49]. The temperature data was obtained from the National Climatic Data Center. It is averaged over the Eastern US (east of 100° W), and shown by black dots in Fig. 4. In this and the following real data experiments, we shall use the default parameters suggested in Section III. The segmentation and state transitions of the time series of the Eastern US temperature over the past century matches the phase shift of the Atlantic Multi-decadal Oscillation (AMO) [4]. As seen from Fig. 4, since the early 20th century,

TABLE II
COMPARISON OF THREE APPROACHES

T	1000			3000			10000		
	3-step	EM	3-step-EM	3-step	EM	3-step-EM	3-step	EM	3-step-EM
MSE of $\hat{\beta}$	0.65	0.97	0.56	0.55	0.81	0.29	0.44	0.53	0.16
	(0.062)	(0.127)	(0.092)	(0.054)	(0.101)	(0.089)	(0.045)	(0.096)	(0.021)
MSE of $\hat{T}$	0.00021	0.95	0.71	0.00020	0.69	0.43	0.00012	0.45	0.23
	(0.000034)	(0.098)	(0.083)	(0.000021)	(0.085)	(0.070)	0.000010	(0.059)	(0.042)
Cost	0.26	9.28	9.28	1.48	19.43	20.08	8.30	50.87	57.68
	(0.002)	(0.805)	(0.800)	(0.019)	(2.068)	(1.455)	(0.718)	(4.715)	(4.590)

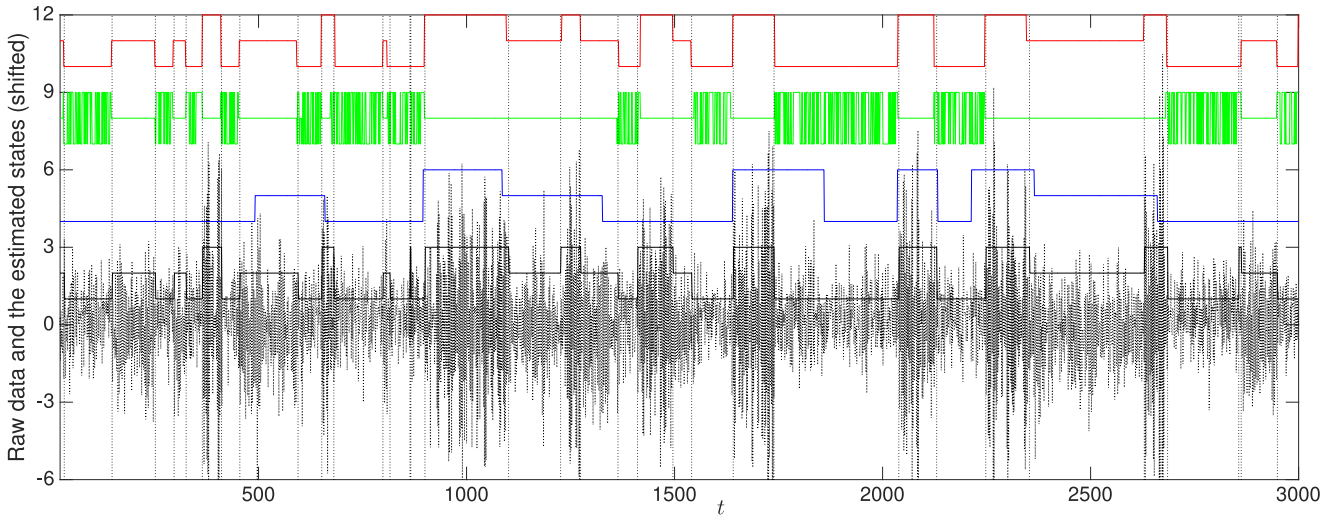


Fig. 3. A figure showing one instance of the time series with $T = 3000, r = 0.9$ (black dash), with the true transition time (black vertical lines), true state $z_t \in \{1, 2, 3\}$ (black line), and the estimated states $\hat{z}_t + 3$ by "3-step" (blue line), $\hat{z}_t + 6$ by "EM" (green line), $\hat{z}_t + 9$ by "3-step-EM" (red line).

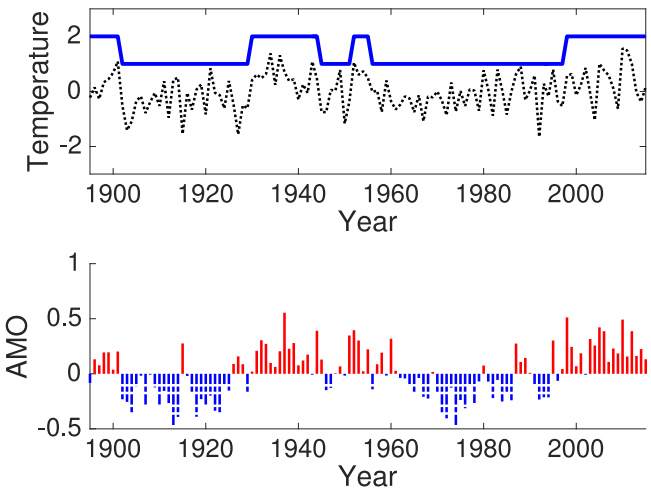


Fig. 4. A figure showing the 1895-2015 summer-time temperature over the Eastern US (unit: °C) and the discovered states.

there has been an alternation between warm phases and cool phases, in synchrony with our discovered states. The dynamic link between AMO and Eastern US climate has previously been reported in the field of environmental science, based upon a global climate model [4], [60]. This validates our conclusion derived from the three-step approach.

2. Chest-mounted accelerometer data: In this experiment, we study the problem of activity recognition using a dataset collected from a wearable accelerometer mounted on the chest, with 52 Hz sampling frequency and 7 labeled activities/states [5]. The 7 states (labeled by 1–7) respectively denote "working at computer", "standing up, walking and going up/down stairs", "standing", "walking", "going up/down stairs", "walking and talking with someone", "talking while standing". The raw data is a 3-dimensional time series recording the acceleration measurements from x, y, and z dimensions. We transform the data to a 1-dimensional series by taking the square norm of acceleration vectors, and then standardize the data. We also subsample the data with rate 52, since the wearer of accelerometer is likely not to frequently change states within a few seconds. By running the three-step approach, we plot the true states (human-labeled), and marked the time steps when we mis-labeled in Fig. 5. Our states have been relabeled to match the true states using a similar technique discussed in Section IV-A.

3. EEG data: In this experiment, we attempt to model EEG data that exhibits structural changes and recurring patterns over time. The data consists of EEG signals recorded from a person over 22 hours with 100 Hz sampling frequency and accompanying hypnograms (expert annotations of sleep stages) [3], [61]. We use $1, \ldots, 6$ to respectively represent the true states '1', '2', '3', '4', 'R', 'W' (which are stored in hypnograms). The raw data was subsampled at rate 10. By running the three-step approach,

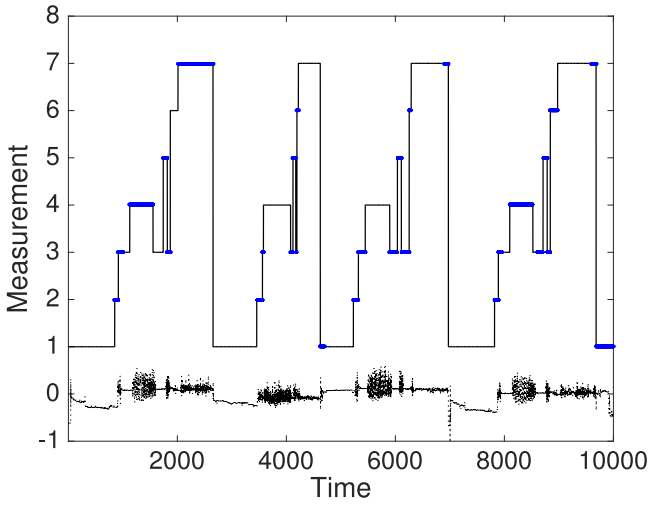


Fig. 5. A figure showing the labeled true states of activities (in black lines), the mis-labeled states (in blue dots), and the pre-processed data (which was rescaled for aesthetic purpose).

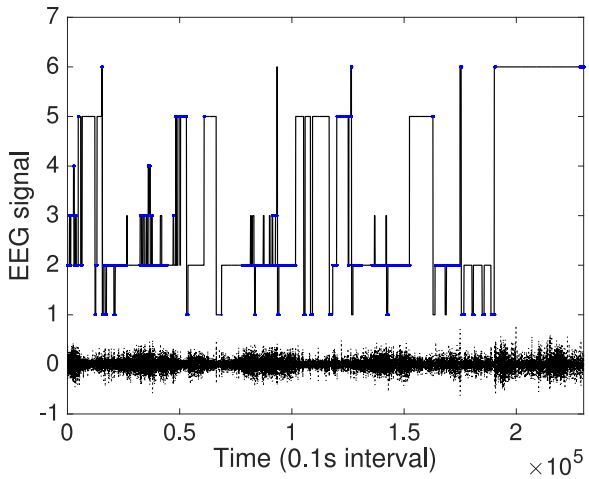


Fig. 6. A figure showing the labeled true states of sleep (in black lines), the mis-labeled states (in blue dots), and the pre-processed data (which was rescaled for aesthetic purpose).

we plot the true and mis-labeled states in Fig. 6. A typical sleep pattern is expected to flow from state 1 to 6, and then to 1 again. Though our discovered state transitions seem reasonable, it is not ideal, perhaps because we were using only one dimensional EEG signals. A possible future work is to extend our three-step procedure to multi-dimensional input, taking into account other signals such as electrooculography (EOG).

V. Conclusion

In this paper, we considered modeling a general dynamic time series which exhibits recurring patterns. By assuming a multi-state autoregressive model with persistent states, we proposed a three-step strategy to efficiently infer the unknown parameters including the number of states. The inferred state sequence can be further modeled to obtain hierarchical information such as Markovity and seasonality.

In this paper, we have assumed that each state is persistent, or the data size between two consecutive change points is large. An interesting future endeavor would be the development of effective strategies for frequent change points. Note that when two change points are very close, it is unrealistic to obtain large sample asymptotics (such as consistent identification of change points). Nevertheless, this may not be an issue when our evaluation metric is merely prediction error. Intuitively, this is possible because if changes are too quick to be discovered, they will have little influence on the average long-term prediction error.

Acknowledgment

The authors thank the Associate Editor and four anonymous reviewers for their constructive comments that have helped to improve the paper.

Appendix

We use $o_p(1)$ and $O_p(1)$ to respectively denote any random variable that converges in probability to zero and that is stochastically bounded.

A. *Proof of Theorem 1*

To prove Theorem 1, we first prove the case of $m = 1$. Let $p_1 = T_1/T, p_2 = (T - T_1)/T$ (which may depend on T). Similar to the definition of X^{T} and $\boldsymbol{\varepsilon}$, we define $X_1, X_2, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2$ by

$$X_1^{\mathrm{T}} = [\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_{T_1}^{\mathrm{T}}], \quad X_2^{\mathrm{T}} = [\boldsymbol{x}_{T_1+1}^{\mathrm{T}}, \ldots, \boldsymbol{x}_T^{\mathrm{T}}],$$
$$\boldsymbol{\varepsilon}_1 = [\varepsilon_1, \ldots, \varepsilon_{T_1}]^{\mathrm{T}}, \quad \boldsymbol{\varepsilon}_2 = [\varepsilon_{T_1+1}, \ldots, \varepsilon_T]^{\mathrm{T}}.$$

It is well known that the least squares estimates of the covariance matrix and noise variance are consistent in the following sense

$$\frac{X^{\mathrm{T}}X}{T}, \frac{X_k^{\mathrm{T}}X_k}{Tp_k} \to_p \Gamma_L, \frac{\hat{e}_{1:T}}{T}, \frac{\hat{e}_{1:T_k}}{Tp_k} \to_p \sigma^2, \tag{9}$$

$k = 1, 2$, and the martingale central limit theorem [47, Appendix 7.5] gives

$$\frac{X^{\mathrm{T}}\boldsymbol{\varepsilon}}{\sqrt{T}}, \frac{X_1^{\mathrm{T}}\boldsymbol{\varepsilon}_1}{\sqrt{Tp_1}}, \frac{X_2^{\mathrm{T}}\boldsymbol{\varepsilon}_2}{\sqrt{Tp_2}} \to_d \mathcal{N}(0, \sigma^2\Gamma_L). \tag{10}$$

It follows that $(\hat{e}_{1:T_1} + \hat{e}_{T_1+1:T})/T \to_p \sigma^2$, and

$$\begin{aligned} g_1 &= T\left(\frac{\hat{e}_{1:T}}{\hat{e}_{1:T_1} + \hat{e}_{T_1+1:T}} - 1\right)\{1 + o(1)\} \\ &= \frac{\hat{e}_{1:T} - \hat{e}_{1:T_1} - \hat{e}_{T_1+1:T}}{(\hat{e}_{1:T_1} + \hat{e}_{T_1+1:T})/T}\{1 + o(1)\} \\ &= \sigma^{-2}(\hat{e}_{1:T} - \hat{e}_{1:T_1} - \hat{e}_{T_1+1:T})\{1 + o_p(1)\}. \end{aligned} \tag{11}$$

Therefore, $g_1 \to_d \chi^2_L$ is equivalent to $g'_1 = \hat{e}_{1:T} - \hat{e}_{1:T_1} - \hat{e}_{T_1+1:T} \to_d \sigma^2\chi^2_L$. We rewrite g'_1 as

$$\begin{aligned} g'_1 &= \sum_{k=1}^{2} \varepsilon_k^{\mathrm{T}} X_k (X_k^{\mathrm{T}} X_k)^{-1} X_k^{\mathrm{T}} \varepsilon_k - \varepsilon^{\mathrm{T}} X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \varepsilon \\ &= \sum_{k=1}^{2} \frac{1}{p_k} \varepsilon_k^{\mathrm{T}} X_k (X^{\mathrm{T}} X)^{-1} X_k^{\mathrm{T}} \varepsilon_k - \varepsilon^{\mathrm{T}} X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \varepsilon \\ &\quad + \sum_{k=1}^{2} \varepsilon_k^{\mathrm{T}} X_k \left\{ (X_k^{\mathrm{T}} X_k)^{-1} - \frac{1}{p_k} (X^{\mathrm{T}} X)^{-1} \right\} X_k^{\mathrm{T}} \varepsilon_k. \end{aligned} \tag{12}$$

From identities (9) and (10), the last two terms are $o_p(1)$ since they can be rewritten as

$$\frac{\varepsilon_k^{\mathrm{T}} X_k}{\sqrt{T p_k}} \left\{ \left(\frac{1}{T p_k} X_k^{\mathrm{T}} X_k \right)^{-1} - \left(\frac{1}{T} X^{\mathrm{T}} X \right)^{-1} \right\} \frac{X_k^{\mathrm{T}} \varepsilon_k}{\sqrt{T p_k}},$$

for $k = 1, 2$. Therefore, using the fact that $p_1 + p_2 = 1$ we obtain

$$\begin{aligned} g'_1 &= \frac{p_2}{p_1} \varepsilon_1^{\mathrm{T}} X_1 (X^{\mathrm{T}} X)^{-1} X_1^{\mathrm{T}} \varepsilon_1 \\ &\quad + \frac{p_1}{p_2} \varepsilon_2^{\mathrm{T}} X_2 (X^{\mathrm{T}} X)^{-1} X_2^{\mathrm{T}} \varepsilon_2 \\ &\quad - \varepsilon_1^{\mathrm{T}} X_1 (X^{\mathrm{T}} X)^{-1} X_2^{\mathrm{T}} \varepsilon_2 \\ &\quad - \varepsilon_2^{\mathrm{T}} X_2 (X^{\mathrm{T}} X)^{-1} X_1^{\mathrm{T}} \varepsilon_1 + o_p(1) \\ &= \left\| \sqrt{\frac{p_2}{T p_1}} \varepsilon_1^{\mathrm{T}} X_1 - \sqrt{\frac{p_1}{T p_2}} \varepsilon_2^{\mathrm{T}} X_2 \right\|^2_{(X^{\mathrm{T}} X/T)^{-1}} \\ &\quad + o_p(1). \end{aligned} \tag{13}$$

Note that $X_1^{\mathrm{T}} \boldsymbol{\varepsilon}_1 / \sqrt{T p_1}$ and $X_2^{\mathrm{T}} \boldsymbol{\varepsilon}_2 / \sqrt{T p_2}$ are asymptotically independent since $\{X_t\}$ is strongly mixing under the assumption on noises. [62]. Then it follows from identities (9), (10), (13) and $p_1 + p_2 = 1$ that $(X^{\mathrm{T}} X/T)^{-1} \to_p \Gamma_L^{-1}$,

$$\sqrt{\frac{p_2}{T p_1}} X_1^{\mathrm{T}} \varepsilon_1 - \sqrt{\frac{p_1}{T p_2}} X_2^{\mathrm{T}} \varepsilon_2 \to_d \mathcal{N}(0, p_1 \Gamma_L + p_2 \Gamma_L),$$

which further implies that $g'_1 \to_d \sigma^2 \chi^2_L$ by Slutsky's theorem.

We now generalize to the case $m \geq 2$. Similar to the above proof, it suffices to prove that $g'_m \to_d \sigma^2 \chi^2_{mL}$.

For any integers $0 \leq a < b < c \leq T$, we let $X_{a:b}^{\mathrm{T}} \triangleq [\boldsymbol{x}_a^{\mathrm{T}}, \ldots, \boldsymbol{x}_b^{\mathrm{T}}]$, $\boldsymbol{\varepsilon}_{a:b} = [\varepsilon_a, \ldots, \varepsilon_b]^{\mathrm{T}}$, and define $g_{a,b,c} \triangleq \hat{e}_{a+1:c} - \hat{e}_{a+1:b} - \hat{e}_{b+1:c}$. Then g'_m may be rewritten as $\hat{e}_{1:T} - \sum_{k=1}^{m+1} \hat{e}_{T_{k-1}+1:T_k} = \sum_{k=1}^{m} g_{T_{k-1},T_k,T}$. Using the proved result for $m = 1$ case, $g_{T_{k-1},T_k,T} \to_d \sigma^2 \chi^2_L$. Therefore, it remains to prove that $g_{T_{k-1},T_k,T}$, $k = 1, \ldots, m$ are asymptotically mutually independent. We prove it by induction. Suppose that $g_{T_{k-1},T_k,T}$, $k = 1, \ldots, \ell$ are asymptotically mutually independent, where $1 \leq \ell \leq m-1$. We now prove the statement for $k = \ell + 1$. From (13) and the related discussions in proving for $m = 1$, we can rewrite $g_{T_{k-1},T_k,T} = \|h_k\|^2_{\Gamma_L^{-1}} + o_p(1)$ for each $k = 1, \ldots, m$, where

$$\begin{aligned} h_k \triangleq \sqrt{p_k} \frac{X^{\mathrm{T}}_{T_{k-1}+1:T_k} \boldsymbol{\varepsilon}_{T_{k-1}+1:T_k}}{\sqrt{T_k - T_{k-1}}} \\ - \sqrt{1-p_k} \frac{X^{\mathrm{T}}_{T_k+1:T} \boldsymbol{\varepsilon}_{T_k+1:T}}{\sqrt{T - T_k}}, \end{aligned}$$

and $p_k \triangleq (T - T_k)/(T - T_{k-1})$. For notational simplicity, we write $h_{\ell+1}$ as

$$h_{\ell+1} = \sqrt{\frac{T - T_{\ell+1}}{T - T_\ell}} Z_1 - \sqrt{\frac{T_{\ell+1} - T_\ell}{T - T_\ell}} Z_2$$

where

$$\begin{aligned} Z_1 &\triangleq \frac{X^{\mathrm{T}}_{T_\ell+1:T_{\ell+1}} \boldsymbol{\varepsilon}_{T_\ell+1:T_{\ell+1}}}{\sqrt{T_{\ell+1} - T_\ell}} \\ Z_2 &\triangleq \frac{X^{\mathrm{T}}_{T_{\ell+1}+1:T} \boldsymbol{\varepsilon}_{T_{\ell+1}+1:T}}{\sqrt{T - T_{\ell+1}}}. \end{aligned}$$

Thus, it remains to prove that $h_{\ell+1}$ is asymptotically independent with $h_1, \ldots, h_\ell$. Define

$$\begin{aligned} h'_{\ell+1} &\triangleq \sqrt{\frac{T_{\ell+1} - T_\ell}{T - T_\ell}} Z_1 + \sqrt{\frac{T - T_{\ell+1}}{T - T_\ell}} Z_2 \\ \Omega &\triangleq \frac{1}{\sqrt{T - T_\ell}} \begin{bmatrix} \sqrt{T - T_{\ell+1}} & -\sqrt{T_{\ell+1} - T_\ell} \\ \sqrt{T_{\ell+1} - T_\ell} & \sqrt{T - T_{\ell+1}} \end{bmatrix} \end{aligned}$$

which satisfies $\Omega^{\mathrm{T}} \Omega = I$. Since $Z_1, Z_2 \to_d \mathcal{N}(0, \sigma^2 \Gamma_L)$ and they are asymptotically independent [62], we obtain

$$\begin{aligned} \begin{bmatrix} h_{\ell+1} \\ h'_{\ell+1} \end{bmatrix} &= (\Omega \otimes I) \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \\ &\to_d \mathcal{N}\big(0, (\Omega \otimes I)(I \otimes \sigma^2 \Gamma_L)(\Omega \otimes I)^{\mathrm{T}}\big) \end{aligned}$$

which is $\mathcal{N}(0, I \otimes \sigma^2 \Gamma_L)$. Therefore, $h_{\ell+1}$ and $h'_{\ell+1}$ are asymptotically independent. Note that $h'_{\ell+1}$ may be rewritten as

$$h'_{\ell+1} = X^{\mathrm{T}}_{T_\ell+1:T} \boldsymbol{\varepsilon}_{T_\ell+1:T} / \sqrt{T - T_\ell},$$

and $h_{\ell+1}$ is a function of $X^{\mathrm{T}}_{T_\ell+1:T_{\ell+1}} \boldsymbol{\varepsilon}_{T_\ell+1:T_{\ell+1}}$, $X^{\mathrm{T}}_{T_{\ell+1}+1:T}$ $\boldsymbol{\varepsilon}_{T_{\ell+1}+1:T}$. Moreover, $h_1, \ldots, h_\ell$ are functions of $X^{\mathrm{T}}_{T_{k-1}+1:T_k}$ $\boldsymbol{\varepsilon}_{T_{k-1}+1:T_k}$, $k = 1, \ldots, \ell$, and $X^{\mathrm{T}}_{T_\ell+1:T} \boldsymbol{\varepsilon}_{T_\ell+1:T}$. It follows that $h_{\ell+1}$ is asymptotically independent with $h_1, \ldots, h_\ell$. This concludes the proof.

Proof of Theorem 2

By similar arguments in the proof of Theorem 1, we have

$$\hat{\beta}_k = \beta_{z_k} + O_p(\tilde{T}^{-1}) \tag{14}$$

$$= \beta_{z_k} + o_p(1), k = 1, \ldots, m+1. \tag{15}$$

where z_k labels the true state. We will construct a proof by contradiction. Suppose that the kth segment is categorized into one of s clusters that contains at least one segment that is not from the same state. Without loss of generality, let this cluster be labeled as 1, and the kth segment be from state 1. Let A_1 denote the set of segments (represented by their labels) from state 1 that are in cluster 1 (by Algorithm 1), and A_1^c the remaining

segments in cluster 1. We use $\#(S)$ to denote the cardinality of a finite set S. By our assumption, $0 < \#(A_1^c) < m+1$. Let $\hat{\beta}_{A_1}$ denote the mean of $\hat{\beta}_i, i \in A_1$, and similarly we define $\hat{\beta}_{A_1^c}$, $\hat{\beta}_{A_1 \cup A_1^c}$. Note that the sum of Euclidean distances within cluster 1 is

$$\begin{aligned}
\ell_s^{(1)} &\triangleq \sum_{i \in A_1 \cup A_1^c} |\hat{\beta}_i - \hat{\beta}_{A_1 \cup A_1^c}|^2 \\
&= \sum_{i \in A_1} |\hat{\beta}_i - \hat{\beta}_{A_1}|^2 + \sum_{i \in A_1^c} |\hat{\beta}_i - \hat{\beta}_{A_1^c}|^2 \\
&\quad + \frac{\#(A_1)\#(A_1^c)}{\#(A_1)+\#(A_1^c)} |\hat{\beta}_{A_1} - \hat{\beta}_{A_1^c}|^2 \qquad (16) \\
&\geq \frac{\min\{\#(A_1), \#(A_1^c)\}}{2} \left\{ \min_{2 \leq i \leq s_0} |\beta_1 - \beta_i|^2 + o_p(1) \right\} \\
&\geq \frac{1}{2} \min_{2 \leq i \leq s_0} |\beta_1 - \beta_i|^2 + o_p(1) \qquad (17)
\end{aligned}$$

where the identity (16) is by direct calculations We first consider the case $s < s_0$. By the pigeon-hole principle, there exist two segments that are categorized into one of s clusters and that are not from the same state. Therefore, by the previous arguments, for all $1 \leq s < s_0$, $\ell_s + sf_T > \ell_s \geq \ell_s^{(1)} > c_0$ for the constant $c_0 \triangleq \min_{1 \leq i,j \leq s_0, i \neq k} |\beta_i - \beta_j|^2/2$ with probability close to one for large T.

We then consider the case $s \geq s_0$. First of all, we prove that each cluster does not contain two segments that are from different states (with probability going to one). If so, by the previous arguments, eventually $\ell_s \geq \ell_s^{(1)} > c_0$. On the other hand, it is a valid configuration that each cluster only contains segments that are from the same states (which is not unique); if so, then it follows from (15) that $\ell_s = o_p(1)$, which is less than c_0 with probability close to one. Therefore, with probability tending to one, two segments that are from different states are not in the same cluster, and $\ell_s + sf_T, s = s_0, \ldots, m$ are less than $\ell_s + sf_T, s = 1, \ldots, s_0 - 1$.

Then, it remains to prove that $\min_{s_0 \leq s \leq m} \ell_s + sf_T$ is achieved at $s = s_0$. In fact, it follows from (14) that $\ell_s - \ell_{s_0} = O_p(\tilde{T}^{-1})$ for $s_0 \leq s \leq m$ which, by assumption (5), is less than f_T with probability tending to one.

REFERENCES

[1] R. J. Hodrick and E. C. Prescott, "Postwar US business cycles: An empirical investigation," *J. Money, Credit, Bank.*, vol. 29, pp. 1–16, 1997.

[2] P. Moran, "The statistical analysis of the Canadian lynx cycle." *Aust. J. Zool.*, vol. 1, no. 3, pp. 291–298, 1953.

[3] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.

[4] R. T. Sutton and D. L. Hodson, "Atlantic ocean forcing of North American and European summer climate," *Science*, vol. 309, no. 5731, pp. 115–118, 2005.

[5] P. Casale, O. Pujol, and P. Radeva, "Personalization and user verification in wearable systems using biometric walking patterns," *Pers. Ubiquitous Comput.*, vol. 16, no. 5, pp. 563–580, 2012.

[6] S. M. Goldfeld and R. E. Quandt, "A Markov model for switching regressions," *J. Econometrics*, vol. 1, no. 1, pp. 3–15, 1973.

[7] J. D. Hamilton, *Time Series Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1994.

[8] J. D. Hamilton, "Regime switching models," in *Macroeconometrics and Time Series Analysis*. New York, NY, USA: Springer, 2010, pp. 202–209.

[9] G. Lindgren, "Markov regime models for mixed distributions and switching regressions," *Scand. J. Statist.*, vol. 5, pp. 81–91, 1978.

[10] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 1970.

[11] A. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1982, vol. 7, pp. 1291–1294.

[12] H. Tong and K. S. Lim, "Threshold autoregression, limit cycles and cyclical data," *J. Roy. Statist. Soc. Ser. B*, vol. 42, pp. 245–292, 1980.

[13] A. V. Vecchia, "Periodic autoregressive-moving average (PARMA) modeling with applications to water resources," *J. Amer. Water Resources Assoc.*, vol. 21, no. 5, pp. 721–730, 1985.

[14] P. H. Franses and R. Paap, *Periodic Time Series Models*. London, U.K.: Oxford Univ. Press, 2004.

[15] R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam, "Structural break estimation for nonstationary time series models," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 223–239, 2006.

[16] C. Levy-leduc and Z. Harchaoui, "Catching change-points with lasso," in *Proc. 21st Int. Conf. Adv. Neural Inf. Process. Syst.*, 2008, pp. 617–624.

[17] H. Ohlsson, L. Ljung, and S. Boyd, "Segmentation of ARX-models using sum-of-norms regularization," *Automatica*, vol. 46, no. 6, pp. 1107–1111, 2010.

[18] D. Angelosante and G. B. Giannakis, "Group lassoing change-points in piecewise-stationary ar signals," in *Proc. 17th Int. Conf. Digit. Signal Process.*, 2011, pp. 1–8.

[19] G. Koop and D. Korobilis, "Forecasting inflation using dynamic model averaging," *Int. Econ. Rev.*, vol. 53, no. 3, pp. 867–886, 2012.

[20] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Bayesian nonparametric methods for learning Markov switching processes," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 43–54, Nov. 2010.

[21] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1569–1585, Apr. 2011.

[22] L. Bauwens, J.-F. Carpantier, and A. Dufays, "Autoregressive moving average infinite hidden Markov-switching models," *J. Bus. Econ. Statist.*, vol. 35, no. 2, pp. 162–182, 2017.

[23] M. Niedzwiecki, *Identification of Time-Varying Processes*. Hoboken, NJ, USA: Wiley, 2000.

[24] G. J. Edelman, M. Roos, A. Bolck, and M. C. Aalders, "Practical implementation of blood stain age estimation using spectroscopy," *IEEE J. Sel. Topics Quantum Electron.*, vol. 22, no. 3, pp. 415–421, May/Jun. 2016.

[25] G. Comert, A. Bezuglov, and M. Cetin, "Adaptive traffic parameter prediction: Effect of number of states and transferability of models," *Transp. Res. C, Emerg. Technol.*, vol. 72, pp. 202–224, 2016.

[26] E. Dogan, "Analyzing the linkage between renewable and non-renewable energy consumption and economic growth by considering structural break in time-series data," *Renewable Energy*, vol. 99, pp. 1126–1136, 2016.

[27] H. Naser, "Estimating and forecasting the real prices of crude oil: A data rich model using a dynamic model averaging (DMA) approach," *Energy Econ.*, vol. 56, pp. 75–87, 2016.

[28] L. Astolfi *et al.*, "Tracking the time-varying cortical connectivity patterns by adaptive multivariate estimators," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 902–913, Mar. 2008.

[29] D. Kugiumtzis, C. Koutlis, A. Tsimpiris, and V. K. Kimiskidis, "Dynamics of epileptiform discharges induced by transcranial magnetic stimulation in genetic generalized epilepsy," *Int. J. Neural Syst.*, vol. 27, no. 7, 2017, Art. no. 1750037.

[30] T. Aste, W. Shaw, and T. Di Matteo, "Correlation structure and dynamics in volatile markets," *New J. Phys.*, vol. 12, no. 8, 2010, Art. no. 085009.

[31] Y. Chen, W. K. Härdle, and U. Pigorsch, "Localized realized volatility modeling," *J. Amer. Statist. Assoc.*, vol. 105, no. 492, pp. 1376–1393, 2010.

[32] G. A. Tularam and R. Reza, "Water exchange traded funds: A study on idiosyncratic risk using Markov switching analysis," *Cogent Econ. Finance*, vol. 4, no. 1, 2016, Art. no. 1139437.

[33] K. A. K. Johansen, "Stability in an era of instability: Regime changes in the market relationship between liquefied petroleum gases, oil and natural gas," in *Proc. 15th IAEE Eur. Conf.: Heading Toward Sustain. Energy Syst.-Evol. or Revolution?*, 2017.

[34] A. Assaf, "The stochastic volatility model, regime switching and value-at-risk (VaR) in international equity markets," *J. Math. Finance*, vol. 7, no. 02, pp. 492–513, 2017.

[35] P. Ciaian, M. Rajcaniova, and d. Kancs, "The economics of bitcoin price formation," *Appl. Econ.*, vol. 48, no. 19, pp. 1799–1815, 2016.
[36] N. O. Jeffries, "A note on testing the number of components in a normal mixture," *Biometrika*, vol. 90, no. 4, pp. 991–994, 2003.
[37] H. Akaike, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, vol. 22, no. 1, pp. 203–217, 1970.
[38] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Inf. Theory*, 1973, pp. 267–281.
[39] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[40] P. A. Naik, P. Shi, and C.-L. Tsai, "Extending the Akaike information criterion to mixture regression models," *J. Amer. Statist. Assoc.*, vol. 102, no. 477, pp. 244–254, 2007.
[41] F. K. Hui, D. I. Warton, and S. D. Foster, "Order selection in finite mixture models: Complete or observed likelihood information criteria?" *Biometrika*, vol. 102, pp. 724–730, 2015.
[42] C. Keribin, "Consistent estimation of the order of mixture models," *Sankhyā: Indian J. Statist., Ser. A*, vol. 62, pp. 49–66, 2000.
[43] J. Ding, V. Tarokh, and Y. Yang, "Bridging AIC and BIC: A new criterion for autoregression," *IEEE Trans. Inf. Theory*, preprint, doi: 10.1109/TIT.2017.2717599.
[44] A. Wald, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Stat.*, vol. 20, no. 4, pp. 595–601, 1949.
[45] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. New York, NY, USA: Springer, 2013.
[46] T. W. Anderson, *The Statistical Analysis of Time Series*. Hoboken, NJ, USA: Wiley, 1971.
[47] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, vol. 734. Hoboken, NJ, USA: Wiley, 2011.
[48] C. S. Wong and W. K. Li, "On a mixture autoregressive model," *J. Roy. Statist. Soc. Ser. B*, vol. 62, no. 1, pp. 95–115, Sep. 2000.
[49] J. Ding, Y. Xiang, L. Shen, and V. Tarokh, "Multiple change point analysis: Fast implementation and strong consistency," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4495–4510, Sep. 2017.
[50] L. Vostrikova, "Detection disorder in multidimensional random processes," *Soviet Math. Doklady*, vol. 24, pp. 55–59, 1981.
[51] A. Scott and M. Knott, "A cluster analysis method for grouping means in the analysis of variance," *Biometrics*, vol. 30, pp. 507–512, 1974.
[52] I. E. Auger and C. E. Lawrence, "Algorithms for the optimal identification of segment neighborhoods," *Bull. Math. Biol.*, vol. 51, no. 1, pp. 39–54, 1989.
[53] Y.-C. Yao, "Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches," *Ann. Statist.*, vol. 12, pp. 1434–1447, 1984.
[54] B. Jackson *et al.*, "An algorithm for optimal partitioning of data on an interval," *IEEE Signal Process. Lett.*, vol. 12, no. 2, pp. 105–108, Feb. 2005.
[55] R. Killick, P. Fearnhead, and I. Eckley, "Optimal detection of changepoints with a linear computational cost," *J. Amer. Statist. Assoc.*, vol. 107, no. 500, pp. 1590–1598, 2012.
[56] J. Ding, M. Noshad, and V. Tarokh, "Data-driven learning of the number of states in multi-state autoregressive models," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput.*, 2015, pp. 418–425.
[57] T. J. Tjalkens, Y. M. Shtarkov, and F. M. Willems, "Sequential weighting algorithms for multi-alphabet sources," in *Proc. 6th Joint Swedish-Russian Int. Workshop Inf. Theory*, 1993, pp. 230–234.
[58] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653–664, May 1995.
[59] G. McLachlan and D. Peel, *Finite Mixture Models*. Hoboken, NJ, USA: Wiley, 2004.
[60] R. T. Sutton and D. L. Hodson, "Climate response to basin-scale warming and cooling of the north atlantic ocean," *J. Climate*, vol. 20, no. 5, pp. 891–907, 2007.
[61] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. E215–220, 2000.
[62] K. B. Athreya and S. G. Pantula, "A note on strong mixing of ARMA processes," *Statist. Probab. Lett.*, vol. 4, no. 4, pp. 187–190, 1986.

Jie Ding, photograph and biography not available at the time of publication.

Shahin Shahrampour, photograph and biography not available at the time of publication.

Kathryn Heal, photograph and biography not available at the time of publication.

Vahid Tarokh, photograph and biography not available at the time of publication.